Final Exam

This is a partial exam of the Data Science course. There are 25 multiple-choice questions. All the questions have the same weight. There is only one correct or best answer for each question. Below are the general instructions:

- You will have a maximum of 2 hours to complete the exam. Students who have applied for special accommodation receive 30 minutes extra time.
- You are not allowed to leave the exam room during the first 30 minutes and last 15 minutes of the normal exam schedule.
- You can use the Jupyter Notebook to perform computations.
- You can use the "?" feature in the Jupyter Notebook to check the documentation of functions.
- You cannot use any reference materials (e.g., smartphone, books) except your one-page double-sided A4 cheatsheet.
- Read each question and all options very carefully.
- Double-check your answers before submission. We suggest answering all questions. We also suggest that you prioritize the questions that you are confident in and then strategically guess the answers for others that you are not confident in.
- Avoid spending too much time on a difficult question. It is better to complete other questions first and then go back to finish the difficult ones.
- You must keep the exam content confidential and are not allowed to copy or distribute the content in any form.

We have defined several terms in the table below. You can refer back to them when the questions mention these terms. The feature (i.e., the input) is x. The ground truth (i.e., the true output) is y. The predicted output is z.

Term	Equation
Identity activation function $f(x)$	$f(x) = x$
Sigmoid activation function $f(x)$	$f(x) = 1/(1 + e^{-x})$
Tanh activation function $f(x)$	$f(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x})$
Hinge loss function $f(y, z)$	$f(y, z) = \max(1 - yz, 0)$
Squared error loss function $f(y, z)$	$f(y, z) = (y - z)^2$
Perceptron loss function $f(y, z)$	$f(y, z) = \max(-yz, 0)$
Logistic loss function $f(y, z)$	$f(y, z) = \log(1 + e^{-yz})$

Binary cross-entropy loss function $f(y, z)$	$f(y, z) = (1 - y) \log_2(1 - z) - y \log_2(z)$
Softmax activation function $f(x)$, where x_i means the i^{th} element in array x, and n means the total number of elements in array x	$f(x_i) = e^{x_i} / \sum_{j=1}^{n} e^{x_j}$
Entropy H for a coin with two sides (one side has probability p_1, and another side has probability p_2)	$H = p_1 \log_2(1/p_1) + p_2 \log_2(1/p_2)$

Multimodal Data Processing Questions

1.0p **1** Which of the following statements about multimodal data processing is **TRUE**?

- ◯ Contrastive Learning brings negative pairs closer and pushes positive pairs far apart.
- ◯ Late fusion means that we concatenate the representations of modalities A and B before making the predictions.
- ◯ Self-attention layers use fixed weights to filter information.
- ◯ Transformers use multi-head attention to look at different aspects of the inputs.

Structured Data Processing Questions

1.0p **2** Suppose we flipped a coin 100 times, and we got 0 heads and 100 tails. What is the entropy of the result in this coin-flipping experiment?

- ◯ 0.25
- ◯ 0.5
- ◯ 1
- ◯ 0

1.0p **3** Which of the following statements about the Decision Tree model is **FALSE**? Notice that we refer to the general Decision Tree, not a specific implementation.

- ◯ Decision Tree can be used for classification and regression tasks.

- ◯ Decision Tree can handle continuous features.
- ◯ When splitting nodes, Decision Tree can use the same feature multiple times.
- ◯ Decision Tree can be trained very well without problems using the misclassification error as the node-splitting strategy.

Text Data Processing Questions

1.0p **4** Given two word embedding vectors for "cat" (0.9, -0.2) and "dog" (0.3, 0.8), what is the cosine similarity between these two word embedding vectors (round to two decimal places)?

Hint: To calculate the cosine similarity between two vectors, you first need to compute the dot product of the vectors (i.e., sum of element-wise multiplication), and then divide it by the product (i.e., multiplication) of their magnitudes. The magnitude of a vector $v = (x, y)$ is $sqrt(x^2 + y^2)$, where $sqrt$ means taking the square root. Use the numpy.sqrt function if you need to compute the square root.

- ◯ 0.05
- ◯ 0.55
- ◯ 0.08
- ◯ 0.14

1.0p **5** Which of the following best describes the output of topic modeling using Latent Dirichlet Allocation?

- ◯ A set of clusters, where each cluster represents a group of similar documents.
- ◯ A set of vectors, where each vector is represented as a word in a high-dimensional space.
- ◯ A set of vectors, where each vector is represented as counts over the words in the text corpus.
- ◯ A set of vectors, where each vector is represented as weights over the words in the text corpus.

1.0p **6** Which of the following about text processing is **FALSE**?

- ◯ Stemming cannot always correctly identify the original form of each word.
- ◯ Word embedding represents each word as a data point in a high-dimensional space.
- ◯ POS tagging labels the role of each word in a particular part of speech, such as verb.

- ◯ Lemmatization simply chops word tails to obtain the word's base form.

1.0p **7** Which of the following is the order that best describes the attention mechanism?

- A: Compute the attention distribution using softmax
- B: Compute attention-weighted sum of encoder output
- C: Get the encoder output values (from the RNN)
- D: Compute attention scores (dot product similarity)
- E: Randomly shuffle the output from the previous step to increase robustness against noise
- F: Transform encoder outputs (dimension reduction)

- ◯ CFDBA
- ◯ CBDEAF
- ◯ CABDEF
- ◯ CFDAB

1.0p **8** Given the following term frequency (TF) table for four words (“apple”, “bike”, “spaceship”, ”tea”) in four documents, which word is the most representative for the first document (with ID 1) according to TF-IDF?

Document ID	TF for “apple”	TF for “bike”	TF for “spaceship”	TF for “tea”
1	2	0	4	8
2	3	2	3	0
3	2	3	0	1
4	4	0	0	0

- ◯ spaceship
- ◯ bike
- ◯ apple
- ◯ tea

1.0p **9** Given an array [3, 1, 0.1], what is the output (round to two decimal places) if we give the array to a softmax layer in the neural network? Notice that the integer or float numbers in the array are separated by commas.

Hint: Use the numpy.exp function if you need to compute the result of an exponential function.

- ◯ 23.91
- ◯ [0.73, 0.24, 0.02]
- ◯ [20.09, 2.72, 1.11]
- ◯ [0.84, 0.11, 0.05]

Image Data Processing Questions

1.0p **10** Given an image, which of the following kernels will blur the image after performing the typical convolution operation (no padding, stride 1) using the kernel? Assume that Python numpy is imported.

- ◯ numpy.array([[0, 0, 0],[2, 0, 0],[0, 0, 0]])
- ◯ numpy.array([[1, 2, 1],[0, 0, 0],[-1, -2, -1]])
- ◯ numpy.array([[0, 0, 0],[0, 2, 0],[0, 0, 0]])
- ◯ numpy.array([[0.0625, 0.125, 0.0625],[0.125, 0.25, 0.125],[0.0625, 0.125, 0.0625]])

1.0p **11** When performing image classification using deep neural networks, we usually randomly rotate input images, crop input images, and change the image colors before feeding the images to the neural network. Which of the following options best describes the purpose of doing this?

- ◯ To reduce the model training time
- ◯ To reduce the size of input (e.g., images)
- ◯ To perform feature selection
- ◯ To combat overfitting by increasing dataset diversity

1.0p **12** Given the following image with size 4 (both width and height), what is the output after performing a max pooling operation with a 2x2 filter (i.e., width 2 and height 2) and stride 2?

2	7	1	6
4	8	5	1
0	4	3	2
1	3	1	2

Output after max pooling:

8	8	6
8	8	5
4	4	3

Output after max pooling:

4	8	5	6

Output after max pooling:

7
8
4
3

Output after max pooling:

8	6
4	3

1.0p **13** You are training a convolutional neural network using the stochastic gradient descent optimizer for an image classification task. During training the model, you observed that the loss and the

model performance metrics did not converge (i.e., alternating between some high and low values). Which of the following is the best action to consider in this situation?

- ◯ Change the ReLU activation functions to the sigmoid activation functions instead.
- ◯ Decrease the batch size during the training process.
- ◯ Decrease the size of the neural network and use less data to train the model.
- ◯ Decrease the learning rate after a certain number of epochs.

1.0p **14** Which of the following is the order that best describes the typical procedure for optimizing a model in PyTorch?

- A: Update the parameters of the model
- B: Backpropagate gradients for every parameter
- C: Accumulate the gradient from the previous steps
- D: Get a batch from the data loader object
- E: Obtain the predictions from the model
- F: Calculate the loss based on the difference between predictions and labels

- ◯ DECBAF
- ◯ EDFACB
- ◯ DBAFE
- ◯ DEFBA

Deep Learning Questions

1.0p **15** Which of the following techniques is **NOT** effective in combating overfitting when training a deep neural network?

- ◯ Regularizing the model weights using some criterion.
- ◯ Augmenting the input data randomly on the fly.
- ◯ Randomly dropping the neurons with a probability.
- ◯ Increasing the number of layers in the neural network.

1.0p **16** What is the difference between a feedforward neural network and a recurrent neural network regarding the model architecture?

- A feedforward network uses activation functions such as ReLU or sigmoid to introduce nonlinearity, while a recurrent network uses linear activation functions.
- Neurons in different layers in a feedforward network are fully connected. But a recurrent network only considers local connectivity, which means that neurons in different layers are partially connected.
- Neurons in a feedforward network are organized in layers, and neurons in the same layer are not connected to each other. But a recurrent network has no such restriction and can have interconnectivity between neurons in the same layer.
- In a recurrent network, the outputs from one layer are fed back as inputs, which forms feedback loops. But the computation of a feedforward network only goes forward, and the outputs are not re-used as inputs.

1.0p **17** Which of the following is **NOT** a commonly used loss function for classification?

- Perceptron loss
- Logistic loss
- Cross-entropy loss
- Squared error loss

1.0p **18** Which of the following is the order that best describes the process of a typical gradient descent algorithm in machine learning?

- A: Initialize model parameters with a starting point
- B: Make a step to update model parameters in the opposite direction to the gradient with a learning rate
- C: Compute the gradient of the error or cost function with respect to model parameters
- D: Repeat the previous two steps until convergence
- E: Make a step to update model parameters in the same direction to the gradient with a learning rate
- F: Repeat the previous two steps until the error or cost is zero

- ACEF
- ACECBF
- ACBCED
- ACBD

1.0p **19** What is the main purpose of regularization when performing Gradient Descent?

- To reduce the number of features in the dataset

- ◯ To schedule the learning rate appropriately
- ◯ To speed up the algorithm and save computer memory
- ◯ To prevent the algorithm from overfitting the training data

1.0p **20** Given an input RGB image (3 channels) of size 32 by 32, a convolutional layer with 10 image filters (each filter has size 5 by 5), a stride of 3, and padding of 1. How many trainable parameters are in this convolutional layer, excluding the bias parameters? The “x” symbol in the following options indicates multiplication. For example, “5 x 5” means 25.

- ◯ 5 x 5 x 10 x 1
- ◯ 5 x 5 x 1 x 3
- ◯ 32 x 32 x 10 x 3
- ◯ 5 x 5 x 10 x 3

1.0p **21** Suppose we have a deep feedforward neural network with two layers. The first layer has four artificial neurons, and the second layer has one neuron. We want to use the network to perform binary classification (blue or orange colored dots) on the data that is shown below. The features are the values of the horizontal axis (X) and vertical axis (Y). Which of the following settings can help us achieve our task well?

- ◯ Use the tanh activation function for all neurons and use the squared error loss.
- ◯ Use the identity activation function for all neurons and use the squared error loss.
- ◯ Use the identity activation function for all neurons and use the cross-entropy loss.
- ◯ Use the ReLU activation function for all neurons and use the cross-entropy loss.

1.0p **22** Suppose that we are performing a prediction task using artificial neurons. Which of the following statements about artificial neurons is **TRUE**?

- ◯ Support Vector Machine uses the sigmoid activation function with the perceptron loss function.
- ◯ The perceptron classifier uses the identity activation function with the hinge loss function.
- ◯ Linear regression uses the identity activation function with the binary cross entropy loss function.
- ◯ Logistic regression uses the sigmoid activation function with the binary cross entropy loss function.

Coding Questions

1.0p **23** You are preprocessing text data for a document classification task. Given a pandas dataframe object D with two columns “class” and “tokens”. Each row means a document. The “class” column contains information about the class that the document belongs to, such as “Business”. The “tokens” column contains a list of tokens, such as ['Wall, St.', 'Bears', 'Claw', 'Back', 'Into']. Which of the options below best describes the function of the following code?

```
D.explode("tokens").groupby(["class",
"tokens"]).size().reset_index(name="n").sort_values(["class", "n"], ascending=[False,
True]).groupby("class").head(10)
```

- ◯ Remove rows with duplicate tokens in the “tokens” column. Return a new dataframe with a new column that counts the frequency of each token among all documents per class.
- ◯ Compute the average frequency of each token in each document per class. Return a new dataframe with a new column that shows the 10 tokens with the highest average frequency.
- ◯ Return a new dataframe with a new column that shows the 10 most used words per class, and their frequency.
- ◯ Return a new dataframe with a new column that shows the 10 least used words per class, and their frequency.

1.0p **24** Given a pandas dataframe object D, where column “w” in each row contains a list of words, such as ['Wall, St.', 'of', 'Bear', 'claw', 'at', 'back', 'dog', 'into']. Which of the options below best describes the function of the following code?

```
S = ["bear", "cat", "dog", "koala", "rabbit"]
D["w"].apply(lambda x: [word for word in x if word.lower() in S])
```

- Select only the rows in the dataframe where the word list in the “w” column contains any word in list S.
- Count the frequency of each word in the “w” column of the dataframe, and then return the words that occur less frequently than a threshold defined in list S.
- For each word list in each row in the “w” column of the dataframe, convert the words to lower cases, and remove words that are in list S.
- For each word list in each row in the “w” column of the dataframe, convert the words to lower cases, and only keep the words that are in list S.

1.0p **25** Given a 1-dimensional numpy array A with integers or floats. All values in the array are different. Assume that numpy is imported. What is the following code doing?

```
numpy.argsort(A)
```

- Return the maximum value in the array A.
- Return the sum of all values in the array A.
- Return the sorted array A in ascending order.
- Return the indices that would sort the array A in ascending order.